# Patterns of genome evolution that have accompanied host adaptation in *Salmonella*

Gemma C. Langridge[a,1], Maria Fookes[a], Thomas R. Connor[a,2], Theresa Feltwell[a], Nicholas Feasey[a], Bryony N. Parsons[b,c], Helena M. B. Seth-Smith[a,3,4], Lars Barquist[a], Anna Stedman[a], Tom Humphrey[d], Paul Wigley[b,e], Sarah E. Peters[f], Duncan J. Maskell[f], Jukka Corander[g], Jose A. Chabalgoity[h], Paul Barrow[i], Julian Parkhill[a], Gordon Dougan[a], and Nicholas R. Thomson[a]

[a]Pathogen Genomics, The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; [b]Institute of Infection and Global Health, University of Liverpool, Liverpool L69 7BE, United Kingdom; [c]Institute of Translational Medicine, University of Liverpool, Liverpool L69 3GE, United Kingdom; [d]College of Medicine, Swansea University, Swansea SA2 8PP, United Kingdom; [e]School of Veterinary Science, University of Liverpool, Liverpool L69 3GB, United Kingdom; [f]Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, United Kingdom; [g]Department of Mathematics and Statistics, University of Helsinki, FI-00014 Helsinki, Finland; [h]Departamento de Desarrollo Biotecnológico, Instituto de Higiene, Facultad de Medicina, Universidad de la República, Montevideo, CP 11600, Uruguay; and [i]School of Veterinary Medicine and Science, University of Nottingham, Nottingham LE12 5RD, United Kingdom

Edited by Ralph R. Isberg, Howard Hughes Medical Institute, Tufts University School of Medicine, Boston, MA, and approved November 26, 2014 (received for review August 29, 2014)

**Many bacterial pathogens are specialized, infecting one or few hosts, and this is often associated with more acute disease presentation. Specific genomes show markers of this specialization, which often reflect a balance between gene acquisition and functional gene loss. Within *Salmonella enterica* subspecies *enterica*, a single lineage exists that includes human and animal pathogens adapted to cause infection in different hosts, including *S. enterica* serovar Enteritidis (multiple hosts), *S.* Gallinarum (birds), and *S.* Dublin (cattle). This provides an excellent evolutionary context in which differences between these pathogen genomes can be related to host range. Genome sequences were obtained from ∼60 isolates selected to represent the known diversity of this lineage. Examination and comparison of the clades within the phylogeny of this lineage revealed signs of host restriction as well as evolutionary events that mark a path to host generalism. We have identified the nature and order of events for both evolutionary trajectories. The impact of functional gene loss was predicted based upon position within metabolic pathways and confirmed with phenotyping assays. The structure of *S.* Enteritidis is more complex than previously known, as a second clade of *S.* Enteritidis was revealed that is distinct from those commonly seen to cause disease in humans or animals, and that is more closely related to *S.* Gallinarum. Isolates from this second clade were tested in a chick model of infection and exhibited a reduced colonization phenotype, which we postulate represents an intermediate stage in pathogen–host adaptation.**

*Salmonella* | host adaptation | pseudogene | metabolism

The central importance of horizontal acquisition of mobile genetic elements in the development of virulence in bacteria has been well described. It has frequently been observed that, as pathogens acquire virulence determinants, they become increasingly adapted to a specific host (1, 2). Exquisitely host-restricted pathogens also often exhibit extensive genome decay, through insertion sequence element proliferation, genomic rearrangement, and/or pseudogene formation (1, 3, 4). Investigating mechanisms involved in host adaptation is key to an understanding of pathogen evolution and has directly translatable relevance to the epidemiology and potentially the control of human and zoonotic infectious disease.

By concentrating upon individual pathogenic clades, insights have been obtained into specific adaptations relating to specific hosts; however, comparative analysis is relatively rare. By broadening this approach to examine multiple human and animal pathogens, derived from a single closely related lineage but with differing host specializations, there is an opportunity to understand the fundamental evolutionary processes involved in host adaptation. Lineage-specific changes that have become fixed can then be distinguished from those stochastic changes that differentiate individual isolates.

A single lineage within *Salmonella enterica* presents such an opportunity. *S. enterica* is a leading cause of foodborne gastroenteritis, globally responsible for 80 million cases annually (5). Differentiation of *S. enterica* is largely based upon somatic (O) and flagellar (H) antigens, but it is increasingly being typed by

## Significance

Common features have been observed in the genome sequences of bacterial pathogens that infect few hosts. These "host adaptations" include the acquisition of pathogenicity islands of multiple genes involved in disease, losses of whole genes, and even single mutations that affect gene function. Within *Salmonella enterica* is a natural model system of four pathogens that are each other's closest relatives, including a host-generalist, two host-specialists, and one with strong host associations. With whole-genome sequences, we aimed to improve our understanding of the number, nature, and order of these host adaptation events, shedding light on how human and animal pathogens arose in the past, and potentially allowing us to predict how emerging pathogens will evolve in the future.

Author contributions: G.C.L., T.H., P.W., D.J.M., J.A.C., P.B., J.P., G.D., and N.R.T. designed research; G.C.L., T.F., B.N.P., A.S., and S.E.P. performed research; J.C. contributed new reagents/analytic tools; G.C.L., M.F., T.R.C., N.F., B.N.P., H.M.B.S.-S., L.B., T.H., P.W., S.E.P., D.J.M., J.C., and N.R.T. analyzed data; and G.C.L., M.F., T.R.C., B.N.P., L.B., P.W., J.C., and N.R.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the European Nucleotide Archive (ENA), www.ebi.ac.uk/ena (accession nos. ERS003157–ERS003161, ERS004906, ERS004907, ERS004909–ERS004912, ERS004914, ERS007756, ERS022673–ERS022687, ERS024552, ERS217197, ERS217199, ERS217201, ERS217204, ERS217211, ERS217214, ERS217216, ERS217218, ERS217222, ERS217225, ERS217233, ERS217235, ERS217238, ERS217239, ERS217241, ERS217242, ERS217252, ERS217264, ERS400244, ERS400251, ERS400254, ERS400256–ERS400259, ERS400262, ERS400264, ERS429283, ERS429284, ERS430067, LK931482, and LK9315020), and annotated plasmid assemblies are available from the European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute, www.ebi.ac.uk (accessions nos. HG970000 and HG970001).

See Commentary on page 647.

[1]To whom correspondence should be addressed. Email: gb4@sanger.ac.uk.

[2]Present address: Cardiff School of Biosciences, Cardiff University, Cardiff CF10 3AX, United Kingdom.

[3]Present address: Functional Genomics Center Zürich, Universität Zürich, CH-8057 Zurich, Switzerland.

[4]Present address: Institute for Veterinary Pathology, Vetsuisse Faculty, Universität Zürich, CH-8057 Zurich, Switzerland.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1416707112/-/DCSupplemental.

genomic methods, such as multilocus sequence typing (MLST). Somatic serogrouping and MLST have identified a single lineage with closely related members that exhibit a range of different host specializations (6–9). These include two of the most important *Salmonella* pathogens: *S.* Enteritidis and *S.* Gallinarum. In addition to the contribution of *S.* Enteritidis to human disease, in many countries both of these pathogens are notifiable diseases in poultry farming and egg production. Despite their close phylogenetic relatedness, they exhibit strikingly different host ranges, with *S.* Enteritidis capable of infecting multiple host species, whereas *S.* Gallinarum is restricted to infection in galliforme birds. This lineage also includes *S.* Gallinarum biovar Pullorum (hereafter *S.* Pullorum), also restricted to galliformes, and *S.* Dublin, which is strongly associated with infection of cattle and more rarely that of humans (10). The *Salmonella* pathogens adapted to particular hosts are associated with a much more invasive disease than generalists like *S.* Enteritidis, which tend to cause enteritis.

Fifty-nine isolates of this *Salmonella* lineage were selected to capture the diversity of sequence types, phage types, and geographical and temporal spread available at the time. Through genome sequencing, we generated a phylogeny to act as a framework with which to reconstruct the evolutionary history of the lineage, onto which observed gene loss and acquisition could be plotted.

This dataset provides a compelling record of the degradation of common metabolic pathways during host specialization to date. We have documented the order of events during the evolution of an entire *Salmonella* lineage. To link this specifically to host adaptation, we have tested isolates occupying key positions in the phylogeny in their cognate host to assess the effects of gene degradation on the manner and severity of disease caused.

## Results and Discussion

**Classical *S.* Enteritidis Is Not the Ancestor of the *S.* Gallinarum Biovars.** To establish an accurate phylogeny, we selected isolates of *S.* Enteritidis, *S.* Gallinarum, *S.* Pullorum, and *S.* Dublin representing different MLSTs, phage types (PTs), and geographical and temporal spread to ensure that we sampled a broad diversity in this lineage. In total, genomic DNA from 59 isolates was used to generate whole genomes by Illumina multiplex sequencing (Table S1). Single-nucleotide polymorphisms (SNPs) were detected by mapping sequences against a pan-chromosome and virulence plasmid pseudomolecule (*SI Methods*). The maximum-likelihood phylogenetic tree shown in Fig. 1 was constructed based upon the core regions of the reference *S.* Enteritidis strain P125109 genome.

The phylogeny of these serotypes is largely consistent with previous studies showing that *S.* Dublin represents an independent clade (9, 11). Our data show that there are ~23,000 SNPs differentiating *S.* Dublin from the most recent common ancestor of *S.* Enteritidis and the avian-adapted salmonellae, *S.* Gallinarum and *S.* Pullorum. However, rather than being a single clade itself, *S.* Enteritidis appears to be more structurally complex: isolate 01-00493-2 occupies a position on the tree basal to all other *S.* Enteritidis and *S.* Gallinarum/*S.* Pullorum complex isolates, and a second clade of *S.* Enteritidis appears within the *S.* Gallinarum/*S.* Pullorum complex (Fig. 1).

Statistical clustering supports the two clades of *S.* Enteritidis depicted in Fig. 1; a Bayesian analysis of population structure (BAPS) was used to identify significant differences in polymorphism sharing across strains (12). This initially indicated that members of clade 2 clustered separately from clade 1. A subsequent level of clustering then separated the *S.* Enteritidis 01-00493-2 isolate labeled "ancestral" in Fig. 1 from the rest of clade 2 (Dataset S1). No further substructure in *S.* Enteritidis was found to be significant according to the Bayesian statistical model.

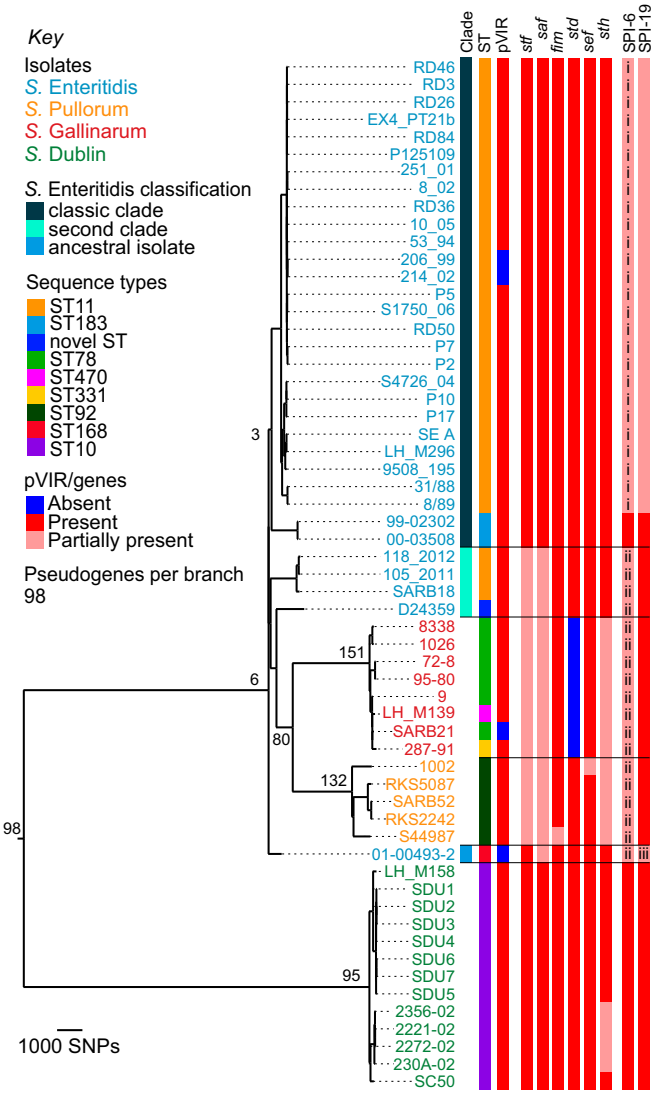


**Fig. 1.** Chromosome-based phylogenetic relationships. Midpoint-rooted maximum-likelihood phylogenetic tree based upon the chromosome. Branch lengths are determined by number of SNPs. Numbers on the tree indicate how many pseudogenes were identified on that branch. Colored strain names show serotype: *S.* Enteritidis, blue; *S.* Pullorum, gold; *S.* Gallinarum, red; *S.* Dublin, green. Metadata columns: Clade, phylogenetic grouping of *S.* Enteritidis isolates; ST, sequence types; pVIR, virulence plasmid; *stf–sth*, fimbriae; SPI-6 and -19, *Salmonella* pathogenicity islands; i, ii, SPI-6 partial hits represent two distinct versions at this locus; iii, SPI-19 partial presence is different from classic *S.* Enteritidis in this strain. Partial presence indicates one or more genes have been lost.

We refer to members of the first clade as "classic" *S.* Enteritidis because this clade includes the most commonly isolated MLST, ST11, and PTs affecting humans and animals. It is of note that both PT11 isolates (99-02302 and 00-03508) occupy a deep branch within the classic clade. PT11s are unusual in that they are rare in human infections and have been reported to be associated with hedgehogs (*Erinaceus europaeus*) (13).

The second clade (Fig. 1, second clade) includes isolates SARB18, 118-2012K-186, 105-2011K-1654, and D24359, which share the same O and H antigen designations as classic *S.* Enteritidis and are closely related at the sequence level. It is this second clade of *S.* Enteritidis from which the ancestors of the *S.* Gallinarum/*S.* Pullorum complex emerged. This phylogeny suggests that *S.* Gallinarum and *S.* Pullorum did not directly

evolve from within the most commonly isolated (classic) clade of *S.* Enteritidis, as was previously postulated (8), and that there is much more substructure in the phylogeny of these important human and animal pathogens than was previously known.

**Chromosomal Macroevolution.** Evolution of pathogenicity by *Salmonella* is strongly associated with the acquisition of mobile genetic elements called *Salmonella* pathogenicity islands (SPIs). Many of these SPIs were acquired very early in the evolution of *S. enterica* (14), and so, perhaps unsurprisingly, their complement was found to be conserved across this entire lineage, with the exception of SPI-6 and SPI-19. Among other functions, these SPIs both encode type VI secretion systems (T6SSs), which are known to promote colonization in the avian gut (15). As previously reported for *S.* Enteritidis PT4 P125109, the majority of the classic *S.* Enteritidis isolates sequenced here harbor degenerate versions of both these SPIs. In these isolates, little more than the *saf* fimbrial operon is maintained on SPI-6, whereas on SPI-19 only 13 out of 33 genes, mainly encoding exported proteins, remain intact. The borders of these deletions are the same for all isolates. We also observed that two isolates of the classic *S.* Enteritidis clade, both PT11, carry intact versions of both SPIs. This indicates that the loss of SPI-6 and SPI-19 is restricted and suggests that it occurred following the divergence of the main classic *S.* Enteritidis group from these strains. That the common ancestor of the entire lineage contained both loci is also supported by their (intact) presence in *S.* Dublin.

SPI-19 is intact in all other isolates sequenced in this study (Fig. 1), and for *S.* Gallinarum and *S.* Pullorum, the retention of a functional SPI-19 is consistent with the T6SS it encodes being important for *S.* Gallinarum survival in chicken macrophages (16). However, the degradation of SPI-6 and loss of the T6SS, seen within the second clade of *S.* Enteritidis and the *S.* Gallinarum/*S.* Pullorum complex, may have only occurred once. Examining the recombination events across the entire lineage revealed that the region surrounding SPI-6 appears to have been independently donated from the *S.* Gallinarum lineage to *S.* Pullorum and isolates in the second clade of *S.* Enteritidis (Fig. 1, SPI-6 version ii; Dataset S1).

Within the chromosome, other clade-specific whole gene losses were related to fimbriae. In *Salmonella*, the loss or inactivation of fimbriae is associated with host adaptation (8, 17). *S.* Enteritidis harbors 13 fimbrial operons, and the variable presence of 6 of these occurred in patterns that matched the phylogeny of isolates in this lineage (Fig. 1). The majority of these differences result from partial- or whole-gene losses in host-restricted *S.* Gallinarum and *S.* Pullorum, but partial losses of the *ssf* and *saf* operons were also observed in the second clade of *S.* Enteritidis.

**Extrachromosomal Macroevolution.** Only four isolates in this study do not carry a version of the *Salmonella* virulence plasmid (pVIR) encoding the virulence-associated *spv* operon (Fig. 1). Using the alignment produced by mapping to a pseudomolecule representing the pan-chromosome combined with the pVIR, we extracted the subset of sequences aligning to the pVIR of *S.* Gallinarum 287-91, which is the largest and most complete virulence plasmid in this lineage. The SNPs identified in this alignment were used to construct a pVIR phylogenetic tree (Fig. S1). The plasmid phylogeny mirrors that of the chromosome for all isolates except for the classic Enteritidis clade, which appears to harbor a pVIR more distantly related to the plasmids carried within this lineage (~1,000 SNPs) than the outgroup, pSLT (~300 SNPs).

The smaller size of the classic *S.* Enteritidis pVIR (~60 kb) relative to that of *S.* Gallinarum (~90 kb) can be largely attributed to major deletions in the *tra* region responsible for conjugal transfer (18). Other differences that contribute to the phylogenetic distance between pVIRs include the presence of genes for a partial K88-like fimbria (19), which was found in all of the *S.* Dublin, *S.* Gallinarum, and *S.* Pullorum pVIRs. A *pefA*-like remnant upstream of SG_p0053 was also conserved across these plasmids, which we believe to be a scar of the *pef* fimbrial operon. This remnant has not been annotated in any published K88-like–containing pVIRs to date but has been randomly mutated in *S.* Gallinarum by transposon mutagenesis. In contrast to the parental *S.* Gallinarum, no morbidity or mortality was observed in chickens inoculated with the resultant mutants (19). This suggests that the *pefA* remnant retains some functionality, although it is not clear what interaction this may have with the K88-like replacement. The *pef* fimbrial operon itself is present only in pVIR-carrying strains of *S.* Enteritidis from the classic clade. Alongside the pVIR phylogeny, this suggests that the ancestral pVIR plasmid of the entire lineage was replaced by a different plasmid before the expansion of the classic *S.* Enteritidis clade. In further support of this, we observe that the second clade and ancestral *S.* Enteritidis all have a pVIR more closely related to those carried by *S.* Gallinarum and *S.* Pullorum than to those found in the classic *S.* Enteritidis clade.

**Table 1. Pseudogenes shared between serovars**

| Serovar | Ent. | Pull. | Gall. | Dub. |
|---|---|---|---|---|
| Ent. | 3 | | | |
| Pull. | 6 (0) | 212 | | |
| Gall. | 6 (0) | 105 (25) | 231 | |
| Dub. | 0 | 18 (10) | 18 (11) | 95 |

Dub., Dublin; Ent., Enteritidis; Gall., Gallinarum; Pull., Pullorum. Total number of pseudogenes shared between serovars. Pseudogenes that have different mutations between serovars are given in parentheses. The 98 ancestral pseudogenes common to all are not included. Only pseudogenes present in all sequenced strains of the serovar are shown here.

**Pseudogene Formation Largely Occurred After Serovar Diversification.** Manual, whole-genome comparisons were used to define the extent to which gene degradation has occurred along the branches of the chromosomal phylogenetic tree. In this context, we defined a pseudogene as a gene harboring a mutation (i.e., premature stop codon, frame shift, truncation, or syntenic deletion) relative to an intact version of that gene. Table 1 summarizes the number of pseudogenes present in each of the respective serovars, and how many shared pseudogenes are due to identical or nonidentical mutations. Full pseudogene sets are listed in Dataset S2. In total, 98 ancestral pseudogenes are common to the entire set of serovars. These are typically fragments of nonfunctional genes and are likely due to stochastic loss. However, pseudogenes that are found in particular regions of the tree are more likely to have resulted from a primary event interrupting metabolic or other pathways, potentially leading to further degradation of related genes.

*S.* Enteritidis is generally regarded as a promiscuous serovar, and our analysis found only one nonancestral pseudogene shared by all classic isolates (Fig. 1). When considered alongside the maintenance of intact fimbriae, replacement of the ancestral plasmid, and loss of both T6SSs, it is apparent that these are markers of adaptation to host generalism in classic *S.* Enteritidis. Fig. S2 depicts the differing routes toward host generalism and adaptation seen across the entire lineage, from a hypothetical shared ancestor.

Previous studies have indicated that the degree of host specificity displayed by particular *Salmonella* serovars loosely correlates with the level of gene degradation (8, 17, 20, 21). In keeping with this observation, host-restricted *S.* Gallinarum and *S.* Pullorum harbor 231 and 212 pseudogenes, respectively, whereas *S.* Dublin, which is associated with but not restricted to cattle, harbors 95 (Table 1).

Significantly, given their shared host restriction, analysis of pseudogene positioning across the tree indicated that over 60%

of the genome degradation found in *S.* Gallinarum (151 of 231) and *S.* Pullorum (132 of 212) occurred after the two diverged. Our recombination analysis revealed that there has been one large recombination event between them, representing a region of around 180 kb (Dataset S1), but this accounts for only six shared pseudogenes, indicating that recombination has not been a significant cause of shared gene loss. However, this region encompasses 165 genes, including four fimbrial operons (*saf*, *sti, stf*, and *stb*) and SPI-6, suggesting that recombination has played some role in these two pathogens converging on the same host niche.

**Loss of Metabolic Capacity.** Functionally, the 98 identified ancestral pseudogenes consist mainly of phage, insertion sequences, and genes of unknown function. This is in stark contrast to the functional categories represented by pseudogenes found toward the tips of the tree (Figs. 1 and 2*A*), where membrane/surface structure and central/intermediary metabolism genes are more commonly inactivated. This latter pattern is broadly consistent for *S.* Gallinarum and *S.* Pullorum, both individually and in their shared 78 pseudogenes, and also in the *S.* Dublin pseudogenes, suggesting that loss of related functions is associated with adaptation to a host organism.

We established that ~15–20% of pseudogenes from each serovar had functional locations according to a database of metabolic pathways predicted to be present in *S.* Enteritidis (Dataset S3). This allowed us to establish whether any metabolic functions were commonly affected (i.e., by the same or different pseudogenes in the same pathway) between serovars. The overlap of the affected pathways is shown in Fig. 2*B* and indicates that many specific pathways or transport reactions are affected in more than one of the host-adapted serovars in our set, and indeed in other host-adapted salmonellae (see below).

One sodium/galactoside transporter has been lost in all of the host-adapted serovars within our set, but degradation of the carbon source D-glucarate is the only full pathway that has been affected (by different mutations) in them all. Interestingly, the human-restricted serovars *S.* Typhi and *S.* Paratyphi A also harbor pseudogenes related to the transport of D-glucarate into the cell, suggesting that the utilization of this carbon source may not be advantageous in causing invasive infection. Two isoenzymes catalyze a key reaction in this pathway, one of which is mutated in *S.* Dublin, whereas both are inactivated in *S.* Gallinarum.

The differing impact of these pseudogenes was supported by a phenotyping screen (Dataset S4), because *S.* Dublin has only suffered a loss of redundancy and therefore remains capable of using D-glucarate, whereas *S.* Gallinarum has lost the function entirely.

There are 13 metabolic pathways and two transport reactions commonly affected in *S.* Gallinarum and *S.* Pullorum (Dataset S3). The majority of these (12 of 15) are due to identical pseudogenes that occurred before their divergence. One of these is in the biosynthesis of the siderophore enterobactin, used to scavenge iron from the environment (22). Iron acquisition in the host is key to many bacterial infections (23), and other iron uptake and transport systems remain intact. This is in contrast to previous findings that human-restricted *S.* Typhi is dependent upon iron uptake through enterobactin via the *fep* genes (24) and therefore suggests that enterobactin is not required in the avian host. However, another of the pathways, putrescine biosynthesis, is also affected in *S.* Typhi, and the same enzyme, ornithine decarboxylase, is mutated in all three. Because the alternative putrescine pathway remains intact, loss of redundancy may have an important role alongside loss of function in host adaptation. This is further supported by the phenotyping screen performed in *S.* Dublin. In accordance with the lower absolute number of pseudogenes, *S.* Dublin has only three pathways affected, and four transport reactions (Dataset S3). Of the four relevant substrates present in the phenotyping screen, *S.* Dublin cannot use one, but remains capable of using three others, indicating that reduction in redundancy is more common than loss of function in this host-adapted serovar.

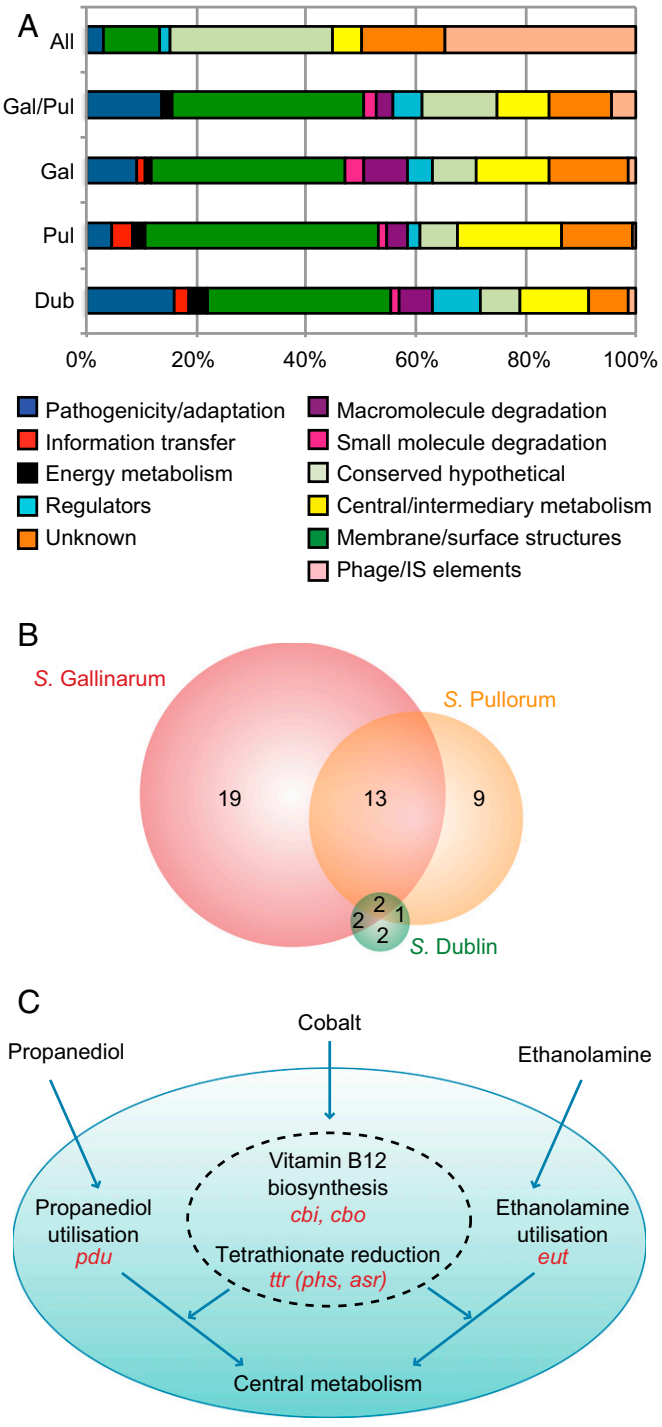


**Fig. 2.** Functions lost through pseudogene formation. (*A*) Functional classification of pseudogene sets. All, pseudogenes shared by all strains in the phylogeny (98); Gal/Pul, shared by *S.* Gallinarum and *S.* Pullorum (80); Gal, Pul, Dub, remaining pseudogenes present in all strains of *S.* Gallinarum (151), *S.* Pullorum (132), and *S.* Dublin (95), respectively. (*B*) Venn diagram showing the distribution of metabolic pathway and transport loss between *S.* Gallinarum, *S.* Pullorum, and *S.* Dublin, irrespective of causative pseudogenes. (*C*) Schematic depicting interconnectivity of pseudogene-affected pathways and transport systems. Processes inside the dotted line only occur anaerobically. Operons involved are shown in red.

Three metabolic pathways are affected in *S.* Gallinarum and *S.* Pullorum due to different pseudogenes that occurred after the two diverged. Two of these functions, allantoin degradation and adenosylcobalamin (vitamin B12) biosynthesis, are also mutated in *S.* Typhi and *S.* Paratyphi A. Given the pseudogene accumulation in these pathways in host-restricted serovars, they emerge as strong contenders for markers of a switch to invasive rather than enteric disease.

In birds, allantoin can be found in the serum and is used as a carbon source during *S.* Enteritidis infection of chickens (25). Inactivation of the genes encoding the regulator *allS* and the degradative enzyme *allD* in *S.* Gallinarum and *S.* Pullorum, respectively, means that neither can use allantoin. Although the relevance of allantoin in mammalian hosts is unknown, pseudogenes relating to allantoin utilization have also been identified in a strain of *S.* Typhimurium belonging to ST313, which causes invasive nontyphoidal salmonellosis in humans (26).

Vitamin B12 is required for the anaerobic degradation of 1,2-propanediol, which uses tetrathionate as an electron acceptor (27). Tetrathionate plays an important role in enteric infection: its production is triggered by the inflammatory response to *S.* Typhimurium in the gut and provides a niche for respiration, in competition with other gut microbes (28). A recent report by Nuccio and Bäumler (21) identified loss of function in central anaerobic metabolism as a key indicator of invasive versus gastrointestinal salmonellae. From plotting pseudogenes onto the chromosomal phylogeny, we know that *S.* Gallinarum and *S.* Pullorum lost the ability to reduce tetrathionate before their divergence, and that their subsequent loss of vitamin B12 synthetic ability occurred independently. The knock-on effect of those initial pseudogenes in tetrathionate reduction is illustrated in Fig. 2*C*, as further pseudogenes have also arisen in other related functions, consistent with Nuccio and Bäumler's findings.

**Second Clade and Ancestral *S.* Enteritidis Display Intermediate Characteristics.** Of the five *S.* Enteritidis isolates outside the classic clade, one (ancestral) was basal to all *S*. Enteritidis and *S*. Gallinarum, whereas the remainder (second clade) were basal to both *S*. Gallinarum and *S.* Pullorum (Fig. 1). To assess whether their position on the phylogenetic tree had any phenotypic consequences, we initially looked at colony morphology and motility of the ancestral isolate (01-00493-2) and a representative isolate from the second clade (SARB18). These indicated that both have intermediate characters between *S.* Enteritidis and *S.* Gallinarum (Fig. S3). One explanation for this could be recombination; multiple separate events have recombined ∼175 kb from *S.* Gallinarum or *S.* Pullorum into the second clade isolates, but not into any from the classic *S.* Enteritidis clade (Dataset S1). Approximately 160 kb originating from *S.* Gallinarum recombined into all four of the second clade isolates, around the SPI-6 locus, resulting in the presence of a different degenerate version of this pathogenicity island compared with that carried by the classic clade isolates. A separate event was responsible for the recombination of ∼85 kb into the ancestral isolate 01-00493-2 from *S.* Gallinarum, also including SPI-6. Unlike classic *S.* Enteritidis, SPI-19 remains intact in the ancestral and second clade isolates of *S.* Enteritidis, as it does in *S.* Gallinarum and *S.* Pullorum (Fig. 1).

Our pseudogene analysis was dependent upon a single reference strain per serovar, and therefore new pseudogenes could not be confirmed in the *S.* Enteritidis isolates outside the classic clade. However, metabolic phenotyping evidence suggests numerous differences are present between the SARB18 second clade and 01-004-93-2 ancestral isolate, with each able to use six or eight different carbon sources, respectively, that could not be used by the other isolate (Dataset S4).

Given the genotypic and phenotypic differences observed, we tested the pathogenesis of these two isolates in comparison with the reference strains of *S.* Enteritidis and *S.* Gallinarum in the natural avian host. The *S.* Gallinarum-infected group showed signs of systemic salmonellosis and reached the humane endpoint of this experimental protocol at 5 d postinfection. The remaining groups showed no signs of ill health and the experiment continued for the full 7 d. At postmortem, the *S.* Gallinarum-infected birds displayed considerable hepatosplenomegaly, white spot lesions on the spleen and discoloration of the liver accompanied by "bronzing" on exposure to air, consistent with fowl typhoid. In contrast, the groups infected with *S.* Enteritidis strains showed mild hepatosplenomegaly consistent with infection by this *Salmonella* serovar. Also as expected, *S*. Gallinarum was the most invasive, indicated by significantly higher colony counts in both the spleen and liver on day 5 ($P < 0.01$) (Fig. 3 *A* and *B*). However, the classic *S*. Enteritidis strain P125109 showed significantly higher levels of colonization (bacteria present in the caeca) compared with both the nonclassic *S.* Enteritidis strains and *S.* Gallinarum ($P < 0.05$) (Fig. 3*C*). This was also borne out by histopathological scoring of the tissue (Table S2). This reduced colonization phenotype from the ancestral and second clade *S.* Enteritidis isolates further suggests that they represent intermediate stages in the evolution of host adaptation.

In conclusion, whole-genome comparisons across an entire *Salmonella* lineage have enabled us to establish the progression of pseudogene formation that has resulted in differently host-adapted pathogens. The strongest signal of metabolic loss was seen in *S.* Gallinarum and *S.* Pullorum, the fully host-restricted

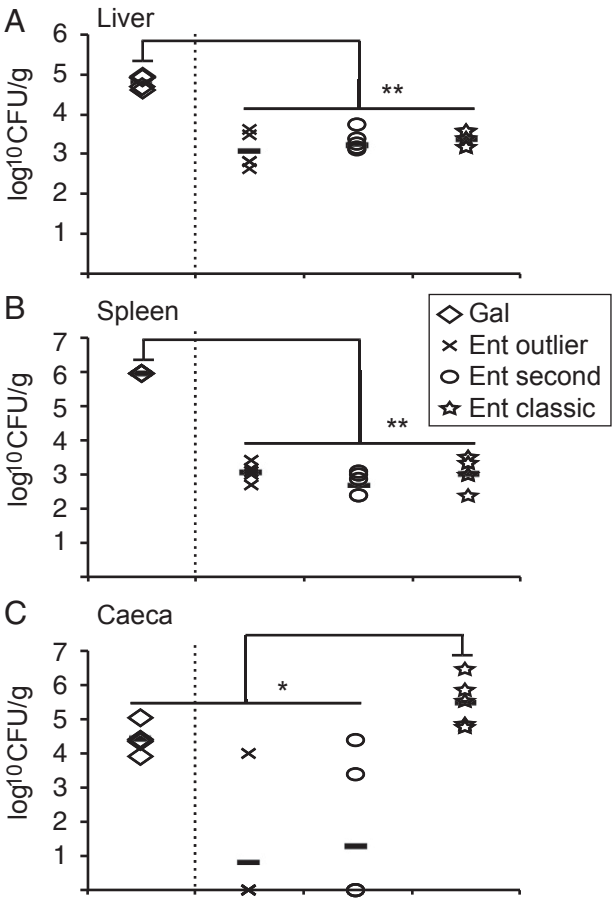


**Fig. 3.** Infection of the avian host by nonclassic *S.* Enteritidis isolates. Invasion of *Salmonella* strains into the (*A*) spleen, (*B*) liver, and (*C*) colonization of chick ceca 7 d (except *S.* Gallinarum 287/91, 5 d, separated by dashed line) postinfection ($n = 5$). Solid lines represent means. $*P < 0.05$, $**P < 0.01$. classic, classic clade isolate; Ent, *S.* Enteritidis; GAL, *S.* Gallinarum; second, second clade isolate.

members of the lineage, consistent with patterns seen in other host-restricted salmonellae. It is therefore plausible that pseudogene formation progresses in a predictable fashion, given a particular host. We also observed the presence of *S.* Enteritidis isolates that fall outside the classic host generalist clade, instead occupying positions basal to the avian-restricted strains. Their metabolic, genomic, and infection phenotypes all suggest that these represent intermediate, but extant, stages in the process of pathogen–host adaptation, demonstrating an unexpected diversity in this serovar.

## Methods

**Sequencing and Phylogenetic Analysis.** All isolates were sequenced on the Illumina platform, with additional 454 sequencing where required. A nonredundant pseudomolecule was used a reference for mapping and to produce genome alignments. The genetic structure of the population was estimated with the software BAPS, version 6.0 (12, 29), and recombination detected using the BratNextGen method (30). Pseudogene identification was performed by manual genome comparison, and metabolic pathway reconstruction was based upon the annotated genome of *S.* Enteritidis, using Pathway Tools software (SRI International). High-throughput phenotype screening was performed using the Biolog system (Biolog). Details are described in *SI Methods*.

**Chick Infection Models.** All experiments were conducted in accordance with United Kingdom legislation governing experimental animals under project licenses PPL 40/3063 and PPL 40/3652 and were approved by the University of Liverpool ethical review process before the award of the license. Four or five chicks were inoculated per experimental group and were killed at 5 or 7 d postinfection. Colony-forming units were quantified from the liver, spleen, and cecal contents, and pathology scoring on hematoxylin-and-eosin–stained sections was performed. Full details are described in *SI Methods*.

**ACKNOWLEDGMENTS.** We thank Mark Achtman and Mark Stevens for kindly providing strains. M.F., T.R.C., H.M.B.S.-S., J.P., and N.R.T. were funded by Wellcome Trust Grant 098051. The chicken infection work by B.N.P., T.H., and P.W. was funded by the Houghton Trust. G.C.L. was funded by Microme, a European Union Framework 7 Programme Collaborative Project, Grant Agreement 222886-2. J.C. was funded by European Research Council Grant 239784 and Academy of Finland Grant 251170.

1. Parkhill J, et al. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* 35(1): 32–40.
2. Rohmer L, et al. (2007) Comparison of *Francisella tularensis* genomes reveals evolutionary events associated with the emergence of human pathogenic strains. *Genome Biol* 8(6):R102.
3. Cole ST, et al. (2001) Massive gene decay in the leprosy bacillus. *Nature* 409(6823): 1007–1011.
4. Moran NA, Plague GR (2004) Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* 14(6):627–633.
5. Majowicz SE, et al.; International Collaboration on Enteric Disease "Burden of Illness" Studies (2010) The global burden of nontyphoidal *Salmonella gastroenteritis*. *Clin Infect Dis* 50(6):882–889.
6. Boyd EF, et al. (1993) *Salmonella* reference collection B (SARB): Strains of 37 serovars of subspecies I. *J Gen Microbiol* 139(Pt 6):1125–1132.
7. Porwollik S, et al. (2005) Differences in gene content between *Salmonella enterica* serovar Enteritidis isolates and comparison to closely related serovars Gallinarum and Dublin. *J Bacteriol* 187(18):6545–6555.
8. Thomson NR, et al. (2008) Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res* 18(10):1624–1637.
9. Achtman M, et al.; *S.* Enterica MLST Study Group (2012) Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog* 8(6):e1002776.
10. Selander RK, et al. (1992) Molecular evolutionary genetics of the cattle-adapted serovar *Salmonella* dublin. *J Bacteriol* 174(11):3587–3592.
11. Olsen JE, Skov M (1994) Genomic lineage of *Salmonella enterica* serovar Dublin. *Vet Microbiol* 40(3-4):271–282.
12. Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* 9(1):539.
13. Nauerby B, Pedersen K, Dietz HH, Madsen M (2000) Comparison of Danish isolates of *Salmonella enterica* serovar enteritidis PT9a and PT11 from hedgehogs (*Erinaceus europaeus*) and humans by plasmid profiling and pulsed-field gel electrophoresis. *J Clin Microbiol* 38(10):3631–3635.
14. Fookes M, et al. (2011) Salmonella bongori provides insights into the evolution of the Salmonellae. *PLoS Pathog* 7(8):e1002191.
15. Blondel CJ, et al. (2010) Contribution of the type VI secretion system encoded in SPI-19 to chicken colonization by *Salmonella enterica* serotypes Gallinarum and Enteritidis. *PLoS One* 5(7):e11724.
16. Blondel CJ, et al. (2013) The type VI secretion system encoded in *Salmonella* pathogenicity island 19 is required for *Salmonella enterica* serotype Gallinarum survival within infected macrophages. *Infect Immun* 81(4):1207–1220.
17. Parkhill J, et al. (2001) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* 413(6858):848–852.
18. Rodríguez-Peña JM, Buisan M, Ibáñez M, Rotger R (1997) Genetic map of the virulence plasmid of *Salmonella enteritidis* and nucleotide sequence of its replicons. *Gene* 188(1):53–61.
19. Rychlik I, Lovell MA, Barrow PA (1998) The presence of genes homologous to the K88 genes faeH and faeI on the virulence plasmid of *Salmonella gallinarum*. *FEMS Microbiol Lett* 159(2):255–260.
20. Chiu C-H, et al. (2005) The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res* 33(5): 1690–1698.
21. Nuccio S-P, Bäumler AJ (2014) Comparative analysis of *Salmonella* genomes identifies a metabolic network for escalating growth in the inflamed gut. *MBio* 5(2):e00929-14.
22. Pollack JR, Ames BN, Neilands JB (1970) Iron transport in *Salmonella typhimurium*: Mutants blocked in the biosynthesis of enterobactin. *J Bacteriol* 104(2):635–639.
23. Skaar EP (2010) The battle for iron between bacterial pathogens and their vertebrate hosts. *PLoS Pathog* 6(8):e1000949.
24. Barquist L, et al. (2013) A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium. *Nucleic Acids Res* 41(8):4549–4564.
25. Dhawi AA, et al. (2011) Adaptation to the chicken intestine in *Salmonella* Enteritidis PT4 studied by transcriptional analysis. *Vet Microbiol* 153(1-2):198–204.
26. Kingsley RA, et al. (2009) Epidemic multiple drug resistant *Salmonella* Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. *Genome Res* 19(12):2279–2287.
27. Price-Carter M, Tingey J, Bobik TA, Roth JR (2001) The alternative electron acceptor tetrathionate supports B12-dependent anaerobic growth of *Salmonella enterica* serovar typhimurium on ethanolamine or 1,2-propanediol. *J Bacteriol* 183(8):2463–2475.
28. Winter SE, et al. (2010) Gut inflammation provides a respiratory electron acceptor for *Salmonella*. *Nature* 467(7314):426–429.
29. Tang J, Hanage WP, Fraser C, Corander J (2009) Identifying currents in the gene pool for bacterial populations using an integrative approach. *PLOS Comput Biol* 5(8): e1000455.
30. Marttinen P, et al. (2012) Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res* 40(1):e6.